Chapter 2

Nonparametric Bayesian Modeling of Multimodal Time Series

Abstract In this chapter, we take a Bayesian nonparametric approach in defining a prior on the hidden Markov model that allows for flexibility in addressing the problem of modeling the complex dynamics during robot manipulation task. At first, considering the underlying dynamics that can be well-modeled as a hidden discrete Markov process, but in which there is uncertainty about the cardinality of the state space. Through the use of the hierarchical Dirichlet process (HDP), one can examine an HMM with an unbounded number of possible states. Subsequently, the sticky HDP-HMM is investigated for allowing more robust learning of the complex dynamics through a learned bias by increasing the probability of self-transitions. Additionally, although the HDP-HMM and its sticky extension are very flexible time series models, they make a strong Markovian assumption that observations are conditionally independent given the discrete HMM state. This assumption is often insufficient for capturing the temporal dependencies of the observations in real data. To address this issue, we consider extensions of the sticky HDP-HMM for learning the switching dynamical processes with switching linear dynamical system. In the later chapters of this book, we will verify the performances in modeling multimodal time series and present the results of robot movement identification, anomaly monitoring, and anomaly diagnose.

2.1 Introduction

Human can perceive the state of the outside world or the objects that need to be operated through the five perspectives based on past experience and internal models, and can recognize the advantages and disadvantages of self to learn and adjust existing experiences and models. However, robots can only understand the outside world through sensing data, so modeling and analysis of multi-modal time series is a difficult and hot research topic to realize robot perception. The multi-modal information during the execution of the robot is a highly complex and dynamic process that incorporates non-linearities such as motion encoders, vision, forces/torques, haptics, and sounds. Modeling based on the traditional hidden Markov model has the problems of uncertainty of the number of recessive states and rapid conversion of recessive states,

and it is difficult to accurately capture the potential modes of multi-modal data. In this chapter, based on the Bayesian basic theory and the theoretical derivation of the HMM model, combined with the inherent characteristics of the robot's multi-modal sensor information, it proposes that it is more efficient and complicated in terms of complex dynamic phenomena, uncertainty, and model parameter learning. The robust non-parameterized Bayesian model implements a solution in which the number of recessive states of the HMM model is determined by the complexity of the data and a high self-transition expected probability of the recessive state. To a certain extent, the non-parameterized Bayesian model is Research on robot multi-modal information sensing and fusion provides theoretical framework and application guidance.

2.2 Related Works

In neurosciences, studies have shown that the different senses of humans can be closely combined [1], so that they can implement the complex manipulation tasks and environmental perception in daily life. In recent years, inspired by the results of this research in neurosciences, scholars have carried out research on the role of multimodal sensing information fusion for implementing practical manipulation tasks in robotics [2, 3]. Fitzpatrick et al. [4] introduced a method of cross-modal verification, which enhances the robot's perception of multi-modal events by repeating and redundancy of sensing information, so that the robot can learn the underlying correlation between vision and auditory. Wu et al. [5] studied the use of a combination of acceleration and sound measurements to detect structural defects in aircraft parts. Su et al. [6] introduced a multi-modal event detection framework based on the robot's axis hole assembly task, including the robot's end pose and haptic sensing signals. It can be known from aforementioned instances that multimodal fusion refers to the synthesis of sensory data with a certain temporal sequence and spatial relationship from multiple different sensors, modeling at different levels and learning its potential patterns [7]. Then, the essence of multimodal fusion is to realize the modeling and analysis of Multivariate Time Series (MTS). For example, in order to allow the robot can perform flexible operations in an uncertain environment, a force/torque sensor is mounted on the end-effector [8], a tactile sensor and a acoustical sensor are installed on the robot claw, and a visual sensor is placed for monitoring the robot workspace [9–11], as shown in Fig. 2.1. In this way, the multi-sensor fusion method can be used to more comprehensively and accurately monitor environmental characteristics, eliminate signal variability and information uncertainty, and improve system reliability and stability [7].

Time series modeling is pervasive in fields as diverse as speech recognition, robotics, economics, medicine, multimedia, bioinformatics, and system control [12]. In recent years, related modeling methods have emerged endlessly, among which the Hidden Markov Model (HMM) [13] is the most popular method that were used in speech recognition [14–16], computational biology [17, 18], Machine translation [19, 20], cryptanalysis [21, 22], economics [23, 24], and human behavior recogni-

Fig. 2.1 Robotic manipulating platforms integrated with multiple sensors

tion [25, 26], etc. In particular, HMM has been widely used in the field of robotics, mainly including process supervision [27, 28], robot state estimation [29, 30], decision making [31, 32], and learning robot motion primitive transformation [33, 34], etc. In literature [27] indicated that HMM can be used for anomaly monitoring, providing a solution for the follow-up work of this book in anomaly implementation. Although there have been a lot of successful application cases of HMM, this model has two inherent shortcomings [35]: (1) It's difficult to determine the size of the HMM hidden state space (the number of hidden states); (2) HMM follows the Markov chain hypothesis (in a given hidden state, the observation data are distributed independently of each other) and the traditional Expectation Maximization (EM) algorithm is used to infer unknown parameters [36, 37]. This inherent disadvantage greatly weakens the accuracy and efficiency of HMM modeling in time series data.

To address the problems of HMM, non-parameterized Bayesian model [38] and HMM of infinite hidden state space [39] were proposed. Yee Whye The et al. [40] proposed a Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM). HDP-HMM uses the hidden state transition probability distribution and observation distribution of HMM. The methods of introducing hyper-parameters to Bayesian a priori are added separately, so that the number of hidden states can be learned from the complexity of the training data, and is calculated by the Markov Chain Monte Carlo (MCMC) method. The Gibbs Sampling (GS) method infers the unknown parameters of the model [41, 42]. HDP-HMM uses the full Bayesian method to learn the number of hidden states by assuming an infinite state space, it effectively avoids overfitting the data and repeated trial and error model selection. From the process of the model implementation and experimental results, it can be known that HDP-HMM's modeling of hidden states still retains the properties of the original Markov chain, which will cause rapid transitions between hidden states, resulting in a lack of the principle of time continuity. The so-called time continuity means that the occurrence of events can always maintain a certain consistency, which makes it difficult to analyze events in the real world. To address the problems of HDP-HMM, Fox et al. [43–45] proposed a Sticky Hierarchical Dirichlet Process Hidden Markov Model (sHDP-HMM). The sHDP-HMM tackle the rapid transition of hidden states in HMM is to increase the self-transition of hidden states by adding the bias of the probability of each hidden state transition, so as to meet the requirement of time continuity. The experimental

results show that the “sticky” hidden state enhances the consistency modeling of time series data, and can achieve more accurate results than HDP-HMM in speech recognition and multiple speaker recognition experiments [43].

2.3 Theoretical Basis of Bayesian Model

2.3.1 Multimodal Time Series

With the development of multi-modal sensing fusion technology and the diversity of the environment, except the joint encoders, robots often need to add force/torque sensors and tactile sensors to sense the surrounding environment, so that the robot’s observation is often multimodal. In general, a multidimensional time series with dimension D and length T is represented as follows:

$$Y_{T \times D} = [y_1(t), y_2(t), \dots, y_d(t), \dots, y_D(t)], t = 1, 2, \dots, T; 1 \leq d \leq D \quad (2.1)$$

where, t represents a time frame, d represents a variable, and $y_d(t)$ denotes the observation of the d -dimensional variable at the t -th time step. For ease of writing, define the observation vectors for all dimensions at time t as $y_t \in \mathbb{R}^D$. Hence, the Eq. 2.1 is expressed as a matrix with $T \times D$, and the matrix is set as an example of a multidimensional time series, that is,

$$Y_{T \times D} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} y_1(1) & y_2(1) & \dots & y_d(1) & \dots & y_D(1) \\ y_1(2) & y_2(2) & \dots & y_d(2) & \dots & y_D(2) \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ y_1(T) & y_2(T) & \dots & y_d(T) & \dots & y_D(T) \end{bmatrix} \quad (2.2)$$

Each row in Eq. 2.2 represents the observation at time t , and each column represents the data of a signal. This book uses a non-parametric Bayesian hidden Markov model to learn and analyze the multi-dimensional time series [45, 46].

2.3.2 Bayes’ Rules

Assuming two random variables are y and θ respectively, then according to the definition of Bayes’ rule as follows:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \quad (2.3)$$

Where $p(\theta|y)$ is the posterior probability, $p(y|\theta)$ is the likelihood probability, $p(\theta)$ is the prior probability, and $p(y)$ is the standardized constant. From Eq. 2.3, the posterior probability is proportional to the product of the likelihood probability and the prior probability. Bayesian law summarizes the process of Bayesian inference: given the likelihood probability and the standardized constant, the posterior probability can be inferred from the prior probability. The Bayesian time series model mentioned in this book means that the variables in the model are time-dependent, but due to the endless time correlation, the model cannot be established, so it is proposed that the variables require certain constraints are time-independent such as Markov chain constraints. So that two adjacent variables y_{t+1} and y_t can be predicted by Eq. 2.3

$$p(y_{t+1}|y_t) = \frac{p(y_t|y_{t+1})p(y_{t+1})}{p(y_t)} \quad (2.4)$$

Therefore, if y in Eq. 2.3 is expressed as the random variable and θ as parameters of corresponding probability model, where unknown parameter is accompanied by a hyper-parameter, Bayesian theory can be extended to

$$p(\theta|y, \lambda) = \frac{p(y|\theta)p(\theta|\lambda)}{p(y)} \quad (2.5)$$

From Eq. 2.5, we can know that the parameters of the model are estimated by maximizing the posterior probability, that is, the maximum posterior probability estimation method (Maximum a Posteriori, MAP), i.e.

$$\theta_{MAP} = \arg \max_{\theta} p(\theta|y, \lambda) \quad (2.6)$$

If the prior probability $p(\theta|\lambda)$ is a constant, then the maximum posterior estimate is equivalent to the Maximum Likelihood (ML), that is,

$$\theta_{ML} = \arg \max_{\theta} p\{y|\theta\} \quad (2.7)$$

Equations 2.6 and 2.7 are both point estimation methods for model parameters, which will widely used to the later chapters.

2.3.3 Conjugate Prior

As we all know, the posterior distribution and the prior distribution belong to the same type in Bayes' rule such that the prior distribution and the posterior distribution are called conjugate distributions. Where the prior distribution is called the conjugate prior of the likelihood function. Therefore, the used of conjugate priors is often motivated by practical considerations that the parameters θ can't be directly

learned from the observations. Namely, the conjugate priors allow for introducing a computationally tractable mechanism for incorporating new observations into the posterior distribution of the parameters θ . With the Bayesian framework, we are interested in incorporating a prior distribution on the latent model parameter θ for making predication about the new observations. To this end, assuming the associated conditional exist, and given N i.i.d. observations, this predictive likelihood is written by

$$p(y|y_1, \dots, y_N, \lambda) = \int_{\Theta} p(y|\vartheta a) p(\vartheta|y_1, \dots, y_N, \lambda) d\vartheta \quad (2.8)$$

To simplify the computational complexity, we only discusses conjugate priors that require the use of three similarity functions in subsequent chapters. In particular, three common observation model are considered in the hidden Markov model, including Multinomial Distribution, Multivariate Gaussian Distribution as well as Linear Gaussian Regression Model.

• Multinomial Observations

Multinomial Distribution [130, 131] considers a random variable y on a finite sample space $\mathcal{Y} = \{1, \dots, K\}$, its probability mass function can be denoted by $\pi = [\pi_1, \dots, \pi_K]$, and describes the probability of a string of N observations of y taking on values $y_1, \dots, y_n$ by

$$p(y_1, y_2, \dots, y_N|\pi) = \frac{N!}{\prod_k N_k!} \prod_k \pi_k^{N_k}, N_k \triangleq \sum_n \delta(y_n, k) \quad (2.9)$$

Where, the notation $\delta(j, k)$ indicate the discrete Kronecker delta. When $K = 2$, this distribution is referred to as the binomial distribution.

The Dirichlet prior distribution is used for formulating the conjugate prior of the multinomial distribution. A K -dimensional Dirichlet distribution is the conjugate prior for the class of K -dimensional multinomial distribution and is uniquely defined by a set of hyperparameters $\alpha = [\alpha_1, \dots, \alpha_K]$, which has the following form:

$$\begin{aligned} \text{Dir}(\alpha_1, \alpha_2, \dots, \alpha_K) &= p(\pi|\alpha) \\ &= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1}, \alpha_k > 0 \end{aligned} \quad (2.10)$$

Where, $\Gamma(\cdot)$ represents the standard Gamma function. When $K = 2$, this distribution is referred to as the Beta distribution, which can be denoted by $Beta(\alpha_1, \alpha_2)$. The initial value of the distribution in Eq. 2.10 is given by

$$[\pi_i] = \frac{\alpha_i}{\sum_k \alpha_k} \quad (2.11)$$

Then, refer to the conjugacy of the Dirichlet distribution, conditioned on N multinomial observations $y_1, \dots, y_N$, the posterior distribution on π is also Dirichlet that

formulated by

$$p(\pi|y_1, y_2, \dots, y_N, \alpha) \propto p(\pi|\alpha)p(y_1, y_2, \dots, y_N|\pi) \propto \text{Dir}(\alpha_1 + N_1, \dots, \alpha_K + N_K) \quad (2.12)$$

Where, the N_k represents the observed times of observation $y_n = k$. Subsequently, using the normalizing constant of the Dirichlet distribution, and substituting to Eq. 2.8, one can derive the predictive likelihood to be

$$p(y = k|y_1, \dots, y_N, \alpha) = \frac{N_k + \alpha_k}{N + \sum_{k=1}^K \alpha_k} \quad (2.13)$$

• Multivariate Gaussian Observations

Multivariate Gaussian distribution is parameterized by a mean vector μ and covariance matrix Σ , which provides a useful description of continuous-valued random variables that concentrate about a given value and have constrained variability [130]. This distribution over a sample space $y \in \mathcal{Y} = \mathbb{R}^D$ that each observation y is a D -dimensional vector can be derived by

$$\mathcal{N}(y; \mu, \Sigma) = p(y|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (y - \mu)^T \Sigma^{-1} (y - \mu) \right\} \quad (2.14)$$

We conveniently denote this multivariate Gaussian distribution by $\mathcal{N}(\mu, \Sigma)$. Three categories of priors about this distribution in terms of the parameters are considered in this book, including known covariance, known mean, and both the mean and covariance are uncertain.

Known Covariance

For known covariance Σ , we use the normal prior distribution as the conjugate prior on the mean vector μ , and represented by $\mathcal{N}(\mu_0, \Sigma_0)$.

Known Mean

For known mean and only the covariance Σ is uncertain, the conjugate prior is formulated by the inverse-Wishart distribution [130]. The D -dimensional inverse-Wishart distribution with two parameters: covariance parameter Δ and ν degrees of freedom, and given by

$$p(\Sigma|\Delta, \nu) = \frac{|\nu\Delta|^{\frac{\nu}{2}} |\Sigma|^{\frac{\nu+D+1}{2}}}{2^{\frac{\nu D}{2}} \pi^{\frac{D(D-1)}{4}} \prod_{i=1}^k \Gamma(\frac{\nu+1-i}{2})} \exp \left\{ -\frac{1}{2} \text{tr}(\nu\Delta\Sigma^{-1}) \right\} \quad (2.15)$$

Where, $\text{tr}(\cdot)$ denotes the trace of matrix. We write this distribution by $\mathcal{IW}(\nu, \Delta)$, and the first moment is given by calculating the expectation value:

$$E[\Sigma] = \frac{\nu\Delta}{\nu - D - 1} \quad (2.16)$$

Both Covariance and Mean are Uncertain

We use the normal-inverse-Wishart distribution as the conjugate prior when both the mean and covariance are uncertain. This distribution defines a conditionally normal prior on the mean $\mu|\Sigma \sim \mathcal{N}(\vartheta, \Sigma/\kappa)$, and an inverse-Wishart distribution on the covariance, $\Sigma \sim \mathcal{IW}(\nu, \Delta)$. Therefore, the joint prior for uncertain mean and covariance is then given by

$$p(\mu, \Sigma|\kappa, \vartheta, \nu, \Delta) \propto |\Sigma|^{\frac{\nu+D+1}{2}} \exp \left\{ \frac{1}{2} \text{tr}(\nu \Delta \Sigma^{-1}) - \frac{\kappa}{2} (\mu - \vartheta)^T \Sigma^{-1} (\mu - \vartheta) \right\} \quad (2.17)$$

Where, κ represents the degree of trust in the mean generated by this prior distribution, ϑ represents the mean of this prior distribution, ν represents the degree of trust in the variance generated by this prior distribution, and Δ represents the mean of the variance Σ matrix. We will use the notation $\mathcal{NIW}(\kappa, \vartheta, \nu, \Delta)$ to represent the normal-inverse-Wishart distribution with four parameters.

• Multivariate Linear Gaussian Regression Observations

The normal multivariate linear regression model is one in which the observations $y_i \in \mathbb{R}^d$ can be described as a linear combination of a set of known regressors $x_i \in \mathbb{R}^n$ with errors accounted for by additive Gaussian noise, written as

$$y_i = x_{i1}a_1 + \cdots + x_{in}a_n + e_i, \quad e_i \sim \mathcal{N}(0, \Sigma) \quad (2.18)$$

For considering the temporal correlation in a multimodal system, we may combine a set of N observation vectors into a matrix $Y = [y_1 \cdots y_N]$, the regressors also into a matrix $X = [x_1 \cdots x_N]$, and the noise terms into $E = [e_1 \cdots e_N]$ and formulated by

$$Y = AX + E \quad (2.19)$$

where the notation $A = [a_1 \cdots a_N]$ is referred to as the design matrix.

Assuming the noise covariance Σ is known, the conjugate prior on the design matrix A is the matrix-normal distribution [132]. A matrix $A \in \mathbb{R}^{d \times m}$ has a matrix-normal distribution $\mathcal{MN}(A; M, V, K)$ that given by

$$p(A) = \frac{|K|^{\frac{d}{2}}}{|2\pi V|^{\frac{m}{2}}} e^{\frac{1}{2} \text{tr}((A-M)^T V^{-1} (A-M)^K)} \quad (2.20)$$

where, the M is the mean matrix, and V and K^{-1} are related to the covariance along the rows and the columns of A .

Along with Eqs. 2.15 and 2.20, the conjugate prior on the set of parameters A and Σ is the matrix-normal inverse-Wishart prior that places a conditionally matrix-normal prior on matrix A given Σ , that is

$$\begin{cases} \Sigma \sim \mathcal{IW}(\nu, \Delta) \\ A|\Sigma \sim \mathcal{MN}(A; M, \Sigma, K) \end{cases} \quad (2.21)$$

2.4 Hidden Markov Model

HMM has been a workhorse in pattern recognition able to encode probabilistic state-space model. The HMM is a stochastic process where a finite number of latent states have Markovian state transitions. Conditioned on the mode sequence, the model assumes (discrete or continuous) conditionally independent observations given the latent state. Let z_t denote the latent state of the Markov chain at time $t \in T$, and π_j the state-specific transition distribution for state $j \in K$. Given the state z_t , the observation y_t is conditionally independent of observations and states at other time steps. Then, the Markovian structure on the state sequence and the observation are simply described as:

$$p(z_t | z_{t-1}) = \pi_{z_{t-1}}, \quad p(y_t | z_t) = F(\theta_{z_t}) \quad t > 1. \quad (2.22)$$

where, the state at the first time step is distributed according to an initial transition distributions π_0 ; $F(\cdot)$ represents a family of distributions (e.g., the multinomial for discrete data, or the multivariate Gaussian for real-vector-valued data); θ_{z_t} are the state-specific emission parameters and the $z_{1:T}$ is a state path over the hidden state space. Therefore, the resulting joint density for T observations is given by:

$$p(z_{1:T}, y_{1:T}) = \pi_0(z_1) p(y_1 | z_1) \prod_{t=2}^T p(z_t | z_{t-1}) p(y_t | z_t). \quad (2.23)$$

Then, to calculate the log-likelihood value of an observation vector $y_{1:T} = (y_1, y_2, y_3, \dots, y_T)$, we first write the joint distribution with the Markov property $p(y_t | y_{t-1}, \dots, y_1) = p(y_t | y_{t-1})$ and derive the joint probability of observation $y_{1:T}$ as:

$$\begin{aligned} p(y_{1:T}) &= \sum_{t=1}^T p(y_t | y_{1:t-1}) \\ p(y_t | y_{1:t-1}) &= \sum_{k=1}^K p(z_t = k | y_{1:t-1}) p(y_t | z_t) \end{aligned} \quad (2.24)$$

As illustrated in Eq. 2.23, if the state path $z_{1:T}$ is estimated, the maximum probability of the observation sequence can be obtained. However, the true state-path is hidden from the observations. There are different approaches to compute it: (i) the maximum likelihood state at any given moment; however, this approach does not consider uncertainty; (ii) a marginal probabilistic representation over hidden states at each time step. Here, the log-probability at time t is derived by computing the natural logarithm of the sum of exponentials over all hidden states:

$$\log p(y_{1:T}) = \log \sum_{k=1}^K \exp \left(\frac{\alpha_t(k) \cdot \beta_t(k)}{p(y_{1:T})} \right) \quad (2.25)$$

where $\alpha_t(k) = p(y_{1:t}, z_t)$, $\beta_t(k) = p(y_{t+1:T} | z_t)$, represent the forward message passing and backward message respectively in the standard forward-backward algorithm [37, 47].

$$\begin{aligned} p(z_t = k | y_{1:t-1}) &= \sum_j p(z_t | z_{t-1} = j) p(z_{t-1} | y_{1:t-1}) \\ \alpha_t(k) &= p(z_t = k | y_{0:t}) \\ &= \frac{1}{p(y_t | y_{1:t-1})} p(y_t | z_t = k) p(z_t = k | y_{1:t-1}). \end{aligned} \quad (2.26)$$

The denominator represents the probability of a sequence of observations that can be calculated by Eq. 2.24.

2.4.1 Forward-Backward Algorithm

To learn unknown parameters, the forward-backward algorithm [48] provides an efficient message passing procedure on computing node marginals of interest problems including *filtering* $p(z_n | y_1, \dots, y_n)$, *prediction* $p(z_{n+m} | y_1, \dots, y_n)$, and *smoothing* $p(z_n | y_1, \dots, y_N)$, $N > n$, where, z_n denotes the hidden state and y_n represents the observation. According to the implementation of the belief propagation algorithm in Bayesian graphical model, we define a set of forward and backward messages by

$$\begin{aligned} \text{forward} : \quad \alpha_n(z_n) &\triangleq p(y_1, \dots, y_n, z_n) \\ \text{backward} : \quad \beta_n(z_n) &\triangleq p(y_{n+1}, \dots, y_N | z_n) \end{aligned} \quad (2.27)$$

To address the problem of filtering, we formulate the filtering with the forward messages defined in Eq. 2.27. The filtering is simply generated as

$$\text{filtering} : \quad p(z_n | y_1, \dots, y_n) = \frac{\alpha_n(z_n)}{\sum_z \alpha_n(z)} \quad (2.28)$$

To address the problem of prediction, we can utilize the Markov structure for the underlying chain to derive, that is

$$p(z_{n+m} | y_1, \dots, y_n) = \frac{\sum_{z_{n+m-1}} p(z_{n+m} | z_{n+m-1}) \cdots \sum_{z_n} p(z_{n+1} | z_n) \alpha_n(z_n)}{\sum_z \alpha_n(z)} \quad (2.29)$$

From Eq. 2.29, it shows that the prediction is equivalent to propagating the forward message without incorporating the missing observations $y_{n+1}, \dots, y_{n+m}$. However, we utilize both of the forward and backward messages for addressing the problem of smoothing for any m by

$$\begin{aligned}
p(z_n|y_1, \dots, y_N) &= \frac{p(y_1, \dots, y_N|z_n)p(z_n)}{p(y_1, \dots, y_N)} \\
&= \frac{p(y_1, \dots, y_n|z_n)p(y_{n+1}, \dots, y_N|z_n)p(z_n)}{p(y_1, \dots, y_N)} \quad (2.30) \\
&= \frac{\alpha_n(z_n)\beta_n(z_n)}{\sum_z \alpha_m(z)\beta_m(z)}
\end{aligned}$$

Hence, we can referred to the belief propagation algorithm for addressing the problems of filtering, predication, and smoothing of HMM. Among of them, we can use the forward messages for calculating the marginal likelihood of observations.

2.4.2 Viterbi Algorithm

The Viterbi algorithm is proposed for addressing the problem of the most likely state sequence to have generated an observation sequence $y_1, \dots, y_N$ when given a set of HMM parameters, which provides an efficient dynamic programming approach to computing this Maximum A Posteriori (MAP) hidden state sequence by

$$\begin{aligned}
\hat{z} &= \max_{z_1, \dots, z_N} \pi^0(z_1)p(y_1|z_1) \prod_{n=2}^N p(z_n|z_{n-1})p(y_n|z_n) \quad (2.31) \\
&= \min_{z_1, \dots, z_N} \left[-\log \pi^0(z_1) - \log p(y_1|z_1) + \sum_{n=1}^N -\log p(z_n|z_{n-1}) - \log p(y_n|z_n) \right]
\end{aligned}$$

From the above, we can note that choosing the MAP sequence is not necessarily equivalent to choosing the maximum marginal independently at each node by $\hat{z}_n = \max p(z_n|y_1, \dots, y_N)$. The marginal sequence also may not even be a feasible sequence for the HMM.

The Viterbi algorithm can be used to compute the most probable sequence of states in HMM based on the dynamic programming principle that the minimum cost path to $z_n = k$ is equivalent to the minimum cost path to node z_{n-1} plus the cost of a transition from z_{n-1} to $z_n = k$. Therefore, the MAP hidden Markov model state sequence $\hat{z}_1, \dots, \hat{z}_N$ can be computed in the following four steps that originally presented in literature [45]. Here, we first initialize the minimum path sum to state $z_1 = k$ for each $k \in \{1, \dots, K\}$ by

$$n = 1 : \mathcal{S}_1(z_1 = k) = \log \pi^0(z_1 = k) - \log p(y_1|z_k = k) \quad (2.32)$$

For $n = 2, \dots, N$, and for each $k \in \{1, \dots, K\}$, calculate the minimum path sum th state $z_n = k$ by

$$n = \{2, \dots, N\} : \mathcal{S}_n(z_n = k) = -\log p(y_n | z_n = k) + \min_{z_{n-1}} \{\mathcal{S}_{n-1}(z_{n-1}) - \log p(z_n = k | z_{n-1})\} \quad (2.33)$$

and simplifies to

$$z_{n-1}^*(z_n) = \arg \min_{z_{n-1}} \{\mathcal{S}_{n-1}(z_{n-1}) - \log p(z_n = k | z_{n-1})\} \quad (2.34)$$

thus, computing the probability of hidden state sequence given observation sequence by

$$\min_{z_1, \dots, z_N} -\log p(z_1, \dots, z_N | y_1, \dots, y_N) = \mathcal{S}_N(z_N) \quad (2.35)$$

and then set the notation $\hat{z}_N = \arg \min_{z_N} \mathcal{S}_N(z_N)$. Finally, the most likely hidden state sequence given observations can be iteratively calculated by

$$\hat{z}_n = z_n^*(\hat{z}_{n+1}), \quad n \in \{N-1, \dots, 1\} \quad (2.36)$$

We will propose methods that have straightforward connections with the Viterbi algorithm for anomaly monitoring during robot manipulation in Chap. 4.

2.5 Nonparametric Bayesian Hidden Markov Model

Bayesian methods are widely applied to various fields such as biomedicine, clinical trials, social science, and economics. The collected data are more and more complicated with the rapid development of modern science and technology. These complicated data include high/ultrahigh dimensional data, big data, discrete and continuous data. The aim of this section is to introduce the newly developed Bayesian methods, including Bayesian variable selection (e.g., fixed-dimensional data analysis and high/ultrahigh dimensional data analysis), Bayesian influence analysis (e.g., case deletion method and local influence analysis), Bayesian estimation and clustering methods (e.g., fixed dimensional data, high/ultrahigh dimensional data analysis, Bayesian network, and Bayesian clustering for big data), Bayesian hypothesis test including discrete and continuous random variables, variational Bayesian analysis and Bayesian clinical trials including design and dose-finding algorithm, and their applications in various fields.

Moreover, Bayesian nonparametric methods avoid the often restrictive assumptions of parametric models by defining distributions on function spaces, especially, for the prior function in Bayesian formulation. If effectively designed, these methods allow for data-driven posterior inference. The overview of Bayesian nonparametric inference in recent decades, see [49–51], and we briefly describe two aspects of Bayesian nonparametric methods on the hidden Markov model, including the

Dirichlet process and its hierarchical extension, such that formulated the Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM) for modeling multimodal observations in robotics.

2.5.1 Bayesian Nonparametric Priors

HMMs restrictively assumes a fixed model complexity. In process monitoring, latent states can represent robot primitives. It's clear that not all robot skills have the same number of primitives and that even the same skill might have variation when conducted under different conditions. To introduce flexibility into the number of computed hidden states, we leverage priors as probability measures.

HMMs can also be represented though a set of transition probability measures G_j . Probability measures yield strictly positive probabilities and sum to 1. Consider if instead of using a transition distribution on latent states, we use it across emission parameters $\theta_j \in \Theta$. Then,

$$G_j = \sum_{k=1}^K \pi_{jk} \delta_{\theta_k}, \quad (2.37)$$

where δ_{θ_k} is the unit mass for mode k at θ . The emission parameter θ_j , can be evaluated at time index $t - 1$, such that:

$$p(\theta'_t | \theta'_{t-1}) = G_{j_{t-1}}, \quad p(y_t | \theta'_t) = F(\theta'_t) \quad (2.38)$$

So, given θ_j , different probability weights are assigned to possible successor candidates θ_k . We can also assign a prior to the categorical probability measure G_j . The Dirichlet distribution is a natural selection due to conjugacy. Thus, the transition probabilities $\pi_j = [\pi_{j1} \cdots \pi_{jK}]$ are independent draws from a K -dimensional Dirichlet distribution:

$$\begin{aligned} p(\pi_j) &= \text{Dir}(\alpha_1, \dots, \alpha_K), \quad j = 1, \dots, K \\ \sum_k^K \pi_{jk} &= 1 \end{aligned} \quad (2.39)$$

Additionally, emission parameters are drawn from a base measure H such that, $p(\theta_j) = H$. The Dirichlet process (DP) was used as the base measured instead of the Dirichlet distribution. The DP is a distribution over countably infinite probability measures $G_0 : \Theta \Rightarrow R^+$, where $G_0(\theta) \geq 0$ and $\int_{\Theta} G_0(\theta) d\theta = 1$. The DP has a joint Dirichlet distribution:

$$p(G_0(\theta_j), \dots, G_0(\theta_K)) = \text{Dir}(\gamma H(\theta_j), \dots, \gamma H(\theta_K)). \quad (2.40)$$

We can further summarize the probability measure as $p(G_0) = DP(\gamma, H)$ as:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \quad p(\theta_k) = H, \quad (2.41)$$

where γ is the concentration parameter and H is the base measure over parameter space Θ . The weights β_k are sampled via a stick-breaking construction:

$$\beta_k = v_k \prod_{l=1}^{k-1} (1 - v_l), \quad (2.42)$$

where $p(v_k) = \beta(1, \gamma)$. For succinctness, the stick-breaking process is defined as: $p(\beta) = GEM(\gamma)$. The DP is used to define a prior on the set of HMM transition probability measures G_j . However, if each transition measure G_j is an independent draw from $DP(\gamma, H)$, where H is continuous, like a Gaussian distribution, transition measures lead to non-overlapping support. This means that previously seen modes (robotic primitives) cannot be selected again. To deal with this limitation, a Hierarchical Dirichlet Process (HDP) is used. The latter constructs transition measures G_j on the same support points $(\theta_1, \theta_2, \dots, \theta_K)$. This can be done when G_j is only a variation on a global discrete measure G_0 , such that:

$$\begin{aligned} G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k}, \quad p(\beta | \gamma) = GEM(\gamma) \\ p(\theta_k | G_0) &= G_0 \\ G_j &= \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k}, \quad p(\pi_j | \alpha, \beta) = DP(\alpha, \beta). \end{aligned} \quad (2.43)$$

This HDP is used as a prior on the HMM. The implications are a mode complexity learned from the data and a sparse state representation. The $DP(\alpha, \beta)$ distribution encourages modes with similar transition distributions, but does not distinguish between self- and cross-mode transitions. This is problematic for dynamical data. The HDP-HMM yields large posterior probabilities for mode sequences with unrealistically fast dynamics. Emily B. Fox et al. [52] introduced the sticky parameter into the HDP, yielding the sHDP-HMM, thus the transition probability π_j is defined as:

$$p(\pi_j | \alpha, \kappa, \beta) = DP\left(\alpha + \kappa, \frac{\alpha\beta + \kappa\delta_i}{\alpha + \kappa}\right) \quad (2.44)$$

The sticky HDP increases the expected probability of self-transitions by an amount proportional to κ and leads to posterior distributions with smoothly varying dynamics. Finally, priors are placed on the concentration parameters α and γ , and the sticky parameter κ . Latent state creation is influenced by α and γ , while self-transition probabilities are biased by κ . These priors allow to integrate expert knowledge for better modeling than the expectation-maximization (EM) algorithm traditionally used in HMMs.

2.5.2 The sHDP-VAR-HMM

Many complex dynamical phenomena are not adequately described as conditionally independent of the observations given the state z_t . That is, the observations are a noisy linear combination of some finite set of past observations plus additive white noise. In such cases, the dynamical processes can be modeled by an autoregressive (AR) process. Particularly, an order r vector AR process, denoted by $\text{VAR}(r)$, with observations $y_t \in \mathbb{R}^d$. The $\text{VAR}(r)$ system can be considered an extension of the HMM; instead of having conditionally independent observations given the hidden state sequence, the system has conditionally linear dynamics. Figure 4.9 shows a graphical model representation and compares it to the sHDP-HMM described in previous section.

The VAR has simplifying assumptions that make it a practical choice in applications [52]. The switching regime can be combined with the sHDP-HMM from Sect. 2.5.1 to leverage the expressiveness of the VAR system with the ability of non-parametric priors to learn the mode complexity of the model. The $\text{VAR}(r)$ process, with autoregressive order r consists of a latent state $z_t \in \mathbb{R}^n$ with linear dynamics observed through $y_t \in \mathbb{R}^d$. The observations have mode specific coefficients and process noise as:

$$\begin{aligned} p(z_t) &= \pi_{z_{t-1}}, \quad t > 1 \\ y_t &= \sum_{i=1}^r A_i^{z_t} y_{t-i} + e_t(z_t), \quad p(e_t) = \mathcal{N}(0, \Sigma). \end{aligned} \quad (2.45)$$

We see a generative model for a time-series $\{y_1, y_2, \dots, y_T\}$ of observed multimodal data, a matrix of regression coefficients $\mathbf{A}^{(k)} = [A_1^{(k)} \dots A_r^{(k)}] \in \mathbb{R}^{dx(d*r)}$, and a measurement noise Σ , with a symmetric positive-definite covariance matrix. Given the observation data, we are interested in learning the “ r^{th} ” model order, for which we need to infer $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$. We leverage the Bayesian approach through the placement of conjugate priors on both parameters for posterior inference. As the mean and covariance are uncertain, the Matrix-Normal Inverse-Wishart (MNIW) serves as an appropriate prior on the multivariate AR distribution. The MNIW places a conditionally matrix-normal prior on $\mathbf{A}^{(k)}$ given Σ :

$$p(\mathbf{A}^{(k)} \mid \Sigma) = \mathcal{MN}(\mathbf{A}; M, \Sigma, K), \quad (2.46)$$

and an inverse-Wishart prior on Σ :

$$p(\Sigma) = IW(v_0, S_0). \quad (2.47)$$

See [45] Appendix F.1 for a detailed derivation of resulting posterior and definitions for variables M and K which are parameters that define the matrix-normal portion of the MNIW prior.

2.6 Summary

In this chapter, a Bayesian nonparametric approach in defining a prior on the hidden Markov model that allows for flexibility in addressing the problem of modeling the complex dynamics during robot manipulation task is proposed. considering the underlying dynamics that can be well-modeled as a hidden discrete Markov process, but in which there is uncertainty about the cardinality of the state space. We use the hierarchical Dirichlet process (HDP) for examining an HMM with an unbounded number of possible states. Subsequently, the sticky HDP-HMM is investigated for allowing more robust learning of the complex dynamics through a learned bias by increasing the probability of self-transitions. Although the HDP-HMM and its sticky extension are very flexible time series models, they make a strong Markovian assumption that observations are conditionally independent given the discrete HMM state. This assumption is often insufficient for capturing the temporal dependencies of the observations in real data. Consequently, an extension of the sticky HDP-HMM for learning the switching dynamical processes with switching linear dynamical system is investigated. As the basis, the detailed formulations and theoretical process of those models are explained step-by-step in each sections, and widely used in the later chapters of this book.

References

1. Alho K, Kujala T, Paavilainen P, et al. Auditory processing in visual brain areas of the early blind: evidence from event-related potentials. *Electroencephalogr Clin Neurophysiol*. 1993;86(6):418–427.
2. Xu Z, Li S, Zhou X, Wu Y, Cheng T, Huang D. Dynamic neural networks based kinematic control for redundant manipulators with model uncertainties. *Neurocomputing*. 2019;329(1):255–66.
3. Xu Z, Li S, Zhou X, Cheng T. Dynamic neural networks based adaptive admittance control for redundant manipulators with model uncertainties. *Neurocomputing*. 2019;357(1):271–81.
4. Fitzpatrick P, Arsenio A, Torres-Jara ER. Reinforcing robot perception of multi-modal events through repetition and redundancy and repetition and redundancy. *Interact Stud*. 2006;7(2):171–196.
5. Huadong W, Mel S. Correlation of accelerometer and microphone data in the “coin tap test”. *IEEE Trans Instrum Meas*. 2000;49(3):493–7.
6. Su Z, Kroemer O, Loeb GE, et al. Learning to switch between sensorimotor primitives using multimodal haptic signals. In: *International conference on simulation of adaptive behavior*. Springer; 2016. p. 170–182.
7. Kirchner EA, Fairclough SH, Kirchner F. Embedded multimodal interfaces in robotics: applications, future trends, and societal implications. *The handbook of multimodal-multisensor interfaces: language processing, software, commercialization, and emerging directions-volume*. 2019;1(3):523–76.
8. Xu Z, Li S, Zhou X, Yan W, Cheng T, Dan H. Dynamic neural networks for motion-force control of redundant manipulators: an optimization perspective. *IEEE Trans Ind Electron*. Early access; 2020. <https://doi.org/10.1109/TIE.2020.2970635>.
9. Kappler D, Pastor P, Kalakrishnan M, et al. Data-driven online decision making for autonomous manipulation. In: *Robotics: science and systems*; 2015.

10. Pieskä S, Kaarela J, Mäkelä J. Simulation and programming experiences of collaborative robots for small-scale manufacturing. In: The 2nd international symposium on small-scale intelligent manufacturing systems (SIMS); 2018. p. 1-4.
11. Li Y, Yue Y, Xu D, et al. Folding deformable objects using predictive simulation and trajectory optimization. In: IEEE/RSJ international conference on intelligent robots and systems (IROS); 2015. p. 6000–6006.
12. Brockwell PJ, Davis RA, Calder MV. Introduction to time series and forecasting. Berlin: Springer; 2002.
13. Eddy SR. Hidden markov models. *Curr Opin Struct Biol.* 1996;6(3):361–5.
14. Huang XD, Ariki Y, Jack MA. Hidden Markov models for speech recognition; 1990.
15. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE.* 2009;77(2):257–286.
16. Gales M, Young S. The application of hidden Markov models in speech recognition. *Found Trends Signal Process.* 2008;1(3):195–304.
17. Krogh A, Brown M, Mian IS, et al. Hidden Markov models in computational biology: applications to protein modeling. *J Mol Biol.* 1994;235(5):1501–1531.
18. Krogh A, Larsson B, Von Heijne G, et al. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.* 2001;305(3):567–80.
19. Och FJ, Ney H. The alignment template approach to statistical machine translation. *Comput Linguist.* 2004;30(4):417–449.
20. Och FJ, Ney H. A comparison of alignment models for statistical machine translation. In: Proceedings of the 18th conference on computational linguistics-volume 2. Association for Computational Linguistics; 2000. p. 1086–1090.
21. Karlof C, Wagner D. Hidden Markov model cryptanalysis. In: International workshop on cryptographic hardware and embedded systems. Springer; 2003. p. 17–34.
22. Green PJ, Noad R, Smart NP. Further hidden Markov model cryptanalysis. In: International workshop on cryptographic hardware and embedded systems. Springer; 2005. p. 61–74.
23. Bhar R, Hamori S. Hidden Markov models: applications to financial economics. Springer Science & Business Media; 2004.
24. Mamon RS, Elliott RJ. Hidden Markov models in finance. Berlin: Springer; 2007.
25. Anne V, Elliot D, Eddie J, et al. Sensor tracked points and HMM based classifier for human action recognition. *Comput Sci Emerg Res J.* 2016;5:4–8.
26. Ren L, Patrick A, Efros AA, et al. A data-driven approach to quantifying natural human motion. *ACM Trans Graph (TOG) ACM.* 2005;24:1090–1097.
27. Hovland GE, McCarragher BJ. Hidden Markov models as a process monitor in robotic assembly. *Int J Robot Res.* 1998;17(2):153–168.
28. Hussam A, George R. Process monitoring using hidden Markov models. *Qual Reliab Eng Int.* 2014;30(8):1379–87.
29. Fox M, Ghallab M, Infantes G, et al. Robot introspection through learned hidden markov models. *Artif Intell.* 2006;170(2):59–113.
30. Wu H, Lin H, Guan Y, et al. Robot introspection with Bayesian nonparametric vector autoregressive hidden Markov models. In: 2017 IEEE-RAS 17th international conference on humanoid robotics (Humanoids). IEEE; 2017. p. 882–888.
31. Chuk TY, Chan AB, Shimojo S, et al. Mind reading: discovering individual preferences from eye movements using switching hidden Markov models. In: Proceedings of the 38th annual conference of the cognitive science society, CogSci; 2016.
32. Manogaran G, Vijayakumar V, Varatharajan R, et al. Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering. *Wirel Pers Commun.* 2018;102(3):2099–2116.
33. Kroemer O, Van Hoof H, Neumann G, et al. Learning to predict phases of manipulation tasks as hidden states. In: 2014 IEEE international conference on robotics and automation (ICRA). IEEE; 2014. p. 4009–4014.
34. Kroemer O, Daniel C, Neumann G, et al. Towards learning hierarchical skills for multi-phase manipulation tasks. In: 2015 IEEE international conference on robotics and automation (ICRA). IEEE; 2015. p. 1503–1510.

35. Eddy SR. What is a hidden Markov model?. *Nat Biotechnol.* 2004;22(10):1315.
36. McLachlan G, Krishnan T. *The EM algorithm and extensions.* Hoboken: Wiley; 2007.
37. Leroux BG. Maximum-likelihood estimation for hidden Markov models. *Stoch Process Their Appl.* 1992;40(1):127–143.
38. Hjort NL, Holmes C, Müller P, et al. *Bayesian nonparametrics.* Cambridge: Cambridge University Press; 2010.
39. Gael JV, Teh YW, Ghahramani Z. The infinite factorial hidden Markov model. *Adv Neural Inf Process Syst.* 2009:1697–1704
40. Teh YW, Jordan MI. Hierarchical Bayesian nonparametric models with applications. *Bayesian Nonparametrics.* 2010;1:158–207.
41. Carlo CM. Markov chain monte carlo and gibbs sampling. *Lecture notes for EEB;* 2004. p. 581.
42. Brooks S, Gelman A, Jones G, et al. *Handbook of markov chain monte carlo.* Boca Raton: CRC Press; 2011.
43. Fox EB, Sudderth EB, Jordan MI, et al. A sticky HDP-HMM with application to speaker diarization. *Ann Appl Stat.* 2011:1020–1056.
44. Fox EB, Sudderth EB, Jordan MI, et al. The sticky HDP-HMM: bayesian nonparametric hidden Markov models with persistent states; 2007.
45. Fox EB. *Bayesian nonparametric learning of complex dynamical phenomena.* Massachusetts Institute of Technology; 2009.
46. Fox E, Sudderth EB, Jordan MI, et al. Nonparametric Bayesian learning of switching linear dynamical systems. *Adv Neural Inf Process Syst;* 2009:457–464.
47. Yu SZ, Kobayashi H. An efficient forward-backward algorithm for an explicit-duration hidden Markov model. *IEEE Signal Process Lett.* 2003;10(1):11–4.
48. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE.* 1989;77.
49. Müller Peter, Quintana FA. Nonparametric bayesian data analysis. *Stat Sci.* 19(1):95–110.
50. Sudderth EB. *Graphical models for visual object recognition and tracking.* Ph.D. thesis, MIT, Cambridge; 2006.
51. Walker SG, Damien PL, Purushottam W, Smith AFM. Bayesian non parametric inference for random distributions and related functions. *J R Stat Soc, Ser B.* 1999;61(3):485–527.
52. Fox EB, Sudderth EB, Jordan MI, Willsky AS. Bayesian nonparametric methods for learning Markov switching processes. *IEEE Signal Process Mag.* 2010;1827(6):43–54.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.